C3.212/2:4

A UNITED STATES DEPARTMENT OF COMMERCE PUBLICATION

Technical Notes 4

Contents

page

Comparison of the Variances of the Net and Gross Error Rates . . . . . Max A. Bershad 1

Effect of Nonrespondents on Estimates of Sales Trend . . . . . Max A. Bershad 3

A Proposal for Obtaining Variance Computations as a By-Product of Regular Tabulations . . . . . Margaret Gurney 5

The Difference Between the Median of a Difference and the Difference of the Medians . . . . . Nigel F. Nettheim 8

The Application of a Linkage Model to the Chandrasekar-Deming Technique for Estimating Vital Events . . . . . Benjamin J. Tepping 11

The Use of the Plotting Machine in Plotting Scatter Diagrams . . . . . Benjamin J. Tepping 17

U.S. DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS

PENNSYLVANIA STATE UNIVERSITY
FEB 15 1971
DOCUMENTS COLLECTION

U.S. DEPARTMENT OF COMMERCE

MAURICE H. STANS, Secretary
ROCCO C. SICILIANO, Under Secretary
Harold C. Passer, Assistant Secretary for Economic Affairs

BUREAU OF THE CENSUS

George H. Brown, Director

Joseph F. Daly, Associate Director

These notes present fragments from research investigations by Bureau of the Census staff members over the past few years. The individual notes are primarily methodological in nature and relate mainly to the design, analysis, and evaluation of various Bureau sample surveys and censuses. Beyond this relationship, the notes are not, in general, related to one another or to any particular area of research.

Most of the notes were initiated in an attempt to answer questions which arose during the course of other projects and consequently they provide only partial answers to the questions that motivated the original studies. Nevertheless, it is felt that the material will be of interest apart from the context in which it was prepared and to a larger audience then that previously reached.

Issued January 1971

Library of Congress Card No. A68-7430

SUGGESTED CITATION

U.S. Bureau of the Census, Technical Notes No. 4
Washington, D.C. 1971

For sale by the Bureau of the Census, Washington, D.C. 20233, or any of the field offices of the Department of Commerce. Price 25 cents

COMPARISON OF THE VARIANCES OF THE NET AND GROSS ERROR RATES

Max A. Bershad*

The purpose of this note is to have a readily available record of the variance of the gross error rate and the net error rate.

A simple random sample of n elements from a very large population has been taken. For each sample element two independent measurements have been made by the same method of measurement. The measurements are made on a 0, 1 variable. The results of the two measurement processes can be displayed as follows:

		Second Measurement		
First		1	0	
Measurement	1	a	b	a+b
	0	c	d	c+d
		a+c	b+d	n

The expected value of repeated measurements on the same element, j, over trials is denoted by $\bar{X}_{j.}$. The response deviation for the j-th element on the t-th trial is then $(x_{jt} - \bar{X}_{j.})$. Response deviations of different elements are assumed to be uncorrelated.

The proportion of sample elements for which the first measurement is "1" and the second measurement is "0" is b/n and will be denoted by p_b; and the proportion for which the first measurement is "0" and the second is "1" is c/n and will be denoted by p_c. Thus, the gross error rate is estimated by $p_b + p_c$ while the net error rate is estimated by $p_b - p_c$. The variance of the gross error rate, $\sigma^2_{p_b + p_c}$, and the variance of the net error rate, $\sigma^2_{p_b - p_c}$, differ only in the sign of the last term, i.e.,

$$\sigma^2_{p_b+p_c} = \sigma^2_{p_b} + \sigma^2_{p_c} + 2\,\sigma_{p_b p_c}$$

$$\sigma^2_{p_b-p_c} = \sigma^2_{p_b} + \sigma^2_{p_c} - 2\,\sigma_{p_b p_c}$$

Thus, we can write these variances as

$$\sigma^2_{p_b \pm p_c} = \sigma^2_{p_b} + \sigma^2_{p_c} \pm 2\,\sigma_{p_b p_c}$$

$$= 2\,(\sigma^2_{p_b} \pm \sigma_{p_b p_c})$$

because $\sigma^2_{p_b} = \sigma^2_{p_c}$.

Since we are dealing with a 0, 1 variable, we can consider the four variables displayed in the 2 x 2 table to have a multinomial distribution. Thus, cov (b,c) = $\sigma_{bc} = -n\,P_B P_C$, and since we assume simple random sampling,

$$\text{cov}\,(p_b, p_c) = \sigma_{p_b p_c} = \frac{-P_B P_C}{n}.$$

Since the two trials are assumed to be interchangeable, $P_B = P_C$. Thus,

$$\sigma_{p_b p_c} = \frac{-P_B^2}{n}$$

* The final version was prepared after the author's death by Barbara A. Bailar.

U. S. Depository Copy

Also, from the multinomial distribution we find that

$$\sigma^2_{p_b} = \frac{P_B\ (1-P_B)}{n}$$

Thus,

$$\sigma^2_{p_b \pm p_c} = \frac{2}{n}\ [(P_B - P_B^2) \pm (-P_B^2)]$$

To determine $E\ (p_b) = P_B$, we note that $p_b = \frac{1}{n} \sum_j^n x_{jt}\ (1 - x_{jt'})$ where x_{jt} is the measurement on the j-th element on the first trial and $x_{jt'}$ is the measurement on the second trial, each measurement being a "0" or a "1". The conditional expected value of p_b for the given sample over all trials is $\frac{1}{n} \sum_j^n (\bar{X}_{j.})\ (1 - \bar{X}_{j.})$. Then, the expected value over all samples is

$$E\ (p_b) = \frac{1}{N} \sum_j^N (\bar{X}_{j.})\ (1 - \bar{X}_{j.})$$

$$= \frac{1}{N} \sum_j^N \bar{X}_{j.} - \frac{1}{N} \sum_j^N \bar{X}^2_{j.}$$

$$= \bar{X}_{..} - (\sigma_S^2 + \bar{X}^2_{..})$$

where σ_S^2 is the sampling variance defined by

$$\sigma_S^2 = E\ [\bar{X}_{j.} - \bar{X}_{..}]^2 = E\ \bar{X}^2_{j.} - \bar{X}^2_{..}$$

Thus,

$$E\ (p_b) = (\bar{X}_{..} - \bar{X}^2_{..}) - \sigma_S^2$$

$$= (\bar{X}_{..})\ (1 - \bar{X}_{..}) - \sigma_S^2$$

But $(\bar{X}_{..})\ (1 - \bar{X}_{..})$ is the total variance, σ^2, for a 0, 1 variable, and $\sigma^2 = \sigma_R^2 + \sigma_S^2$, where σ_S^2 denotes the sampling variance and $\sigma_R^2 = E\ (x_{jt} - \bar{X}_{j.})^2$ denotes the simple response variance, the expectation being taken over all trials and samples.

Thus,

$$E\ (p_b) = \sigma^2 - \sigma_S^2$$

$$= \sigma_R^2$$

Then,

$$\sigma^2_{p_b \pm p_c} = \frac{2}{n}\ [(\sigma_R^2 - \sigma_R^4) \pm (-\sigma_R^4)]$$

and

$$\sigma^2_{p_b - p_c} = \frac{2}{n}\ [\sigma_R^2]$$

while

$$\sigma^2_{p_b + p_c} = \frac{2}{n}\ [\sigma_R^2 - 2\ \sigma_R^4].$$

Thus, if the conditions of the second paragraph apply, the variance of the gross error rate is the variance of the net error rate less n times the square of the variance of the net error rate. When the simple response variance is small, the variance of the gross and net error rates are almost the same.

EFFECT OF NONRESPONDENTS ON ESTIMATES OF SALES TREND

Max A. Bershad*

The problem below occurs in many forms particularly with respect to non-response and cut-off surveys.

THE PROBLEM

A sales trend $(1+\Delta_1)$ is computed from the reports of a select group (say respondents) that covers a given proportion (p) of the total sales of the universe. What must be the sales trend $(1+\Delta_2)$ for the non-selected group covering the remainder of the universe, if the sales trend $(1+\Delta)$ for the universe is to fall within a given percentage $(\pm\epsilon)$ of the trend for the select group?

THE SOLUTION

$$(1+\Delta_2) = (1+\Delta_1)\left(1 \pm \frac{\epsilon}{q}\right) \qquad \text{where } q=1-p$$

AN ILLUSTRATION

A reporting group with 85 percent coverage of an industry's volume, reports an increase of 25 percent. What is the allowable range for the balance of the industry such that the true trend for the industry shall be within 5 percent of 1.25 (or an increase of 19 percent to 31 percent)?

$$(1+\Delta_2) = 1.25\left(1 \pm \frac{.05}{.15}\right) = 1.25\ (1 \pm .33)$$
$$= 1.25 \pm .42$$

The balance of the industry can therefore have a trend of anywhere from .83 to 1.67. In other words, the balance would have to drop more than 17 percent or increase more than 67 percent before the reporting group would not reflect the true answer within 5 percent.

The following table presents values of $(1 + \Delta_2)$ for

(i) values of $(1+\Delta_1)$, from .70 to 1.50 and
(ii) values of p from 40% to 95% and
(iii) values of ϵ of .5%, 1%, 2%, 5%

*The final version was prepared after the author's death by Michael Berry. In the original draft the title was Algebraic Note.

LIMITS OF TREND OF NON-COOPERATORS - BEYOND WHICH COOPERATORS
WILL NOT REFLECT THE TRUE TREND TO THE INDICATED ACCURACY

Trend of Cooperators	Accuracy (Percent)	Coverage Accounted for by Cooperators						
		40%	50%	60%	70%	80%	90%	95%
.70	.5	.69 to .71	.69 to .71	.69 to .71	.69 to .71	.68 to .72	.66 to .73	.63 to .77
	1	.69 " .71	.69 " .71	.68 " .72	.68 " .72	.66 " .73	.63 " .77	.56 " .84
	2	.68 " .72	.67 " .73	.66 " .73	.65 " .75	.63 " .77	.56 " .84	.42 " .98
	5	.64 " .76	.63 " .77	.61 " .79	.58 " .82	.52 " .87	.35 " 1.05	0 " 1.40
.80	.5	.79 " .81	.79 " .81	.79 " .81	.79 " .81	.78 " .82	.76 " .84	.72 " .88
	1	.79 " .81	.78 " .82	.78 " .82	.77 " .83	.76 " .84	.72 " .88	.64 " .96
	2	.77 " .83	.77 " .83	.76 " .84	.75 " .85	.72 " .88	.64 " .96	.48 " 1.12
	5	.73 " .87	.72 " .88	.70 " .90	.67 " .93	.60 " 1.00	.40 " 1.20	0 " 1.60
.90	.5	.89 " .91	.89 " .91	.89 " .91	.88 " .91	.88 " .92	.85 " .94	.81 " .99
	1	.88 " .91	.88 " .92	.88 " .92	.87 " .93	.85 " .94	.81 " .99	.72 " 1.08
	2	.87 " .93	.86 " .94	.85 " .94	.84 " .96	.81 " .99	.72 " 1.08	.54 " 1.26
	5	.82 " .97	.81 " .99	.79 " 1.01	.75 " 1.05	.67 " 1.12	.45 " 1.35	0 " 1.80
1.00	.5	.99 " 1.01	.99 " 1.01	.99 " 1.01	.98 " 1.02	.97 " 1.02	.95 " 1.05	.90 " 1.10
	1	.98 " 1.02	.98 " 1.02	.97 " 1.02	.97 " 1.03	.95 " 1.05	.90 " 1.10	.80 " 1.20
	2	.97 " 1.03	.96 " 1.04	.95 " 1.05	.93 " 1.07	.90 " 1.10	.80 " 1.20	.60 " 1.40
	5	.92 " 1.08	.90 " 1.10	.87 " 1.12	.83 " 1.17	.75 " 1.25	.50 " 1.50	0 " 2.00
1.10	.5	1.09 " 1.11	1.09 " 1.11	1.09 " 1.11	1.08 " 1.12	1.07 " 1.13	1.04 " 1.16	.99 " 1.21
	1	1.08 " 1.12	1.08 " 1.12	1.07 " 1.13	1.06 " 1.14	1.04 " 1.16	.99 " 1.21	.88 " 1.32
	2	1.06 " 1.14	1.06 " 1.14	1.04 " 1.16	1.03 " 1.17	.99 " 1.21	.88 " 1.32	.66 " 1.54
	5	1.01 " 1.19	.99 " 1.21	.96 " 1.24	.92 " 1.28	.82 " 1.38	.55 " 1.65	0 " 2.20
1.20	.5	1.19 " 1.21	1.19 " 1.21	1.18 " 1.22	1.18 " 1.22	1.17 " 1.23	1.14 " 1.26	1.08 " 1.32
	1	1.18 " 1.22	1.18 " 1.22	1.17 " 1.23	1.16 " 1.24	1.14 " 1.26	1.08 " 1.32	.96 " 1.44
	2	1.16 " 1.24	1.15 " 1.25	1.14 " 1.26	1.12 " 1.28	1.08 " 1.32	.96 " 1.44	.72 " 1.68
	5	1.10 " 1.30	1.08 " 1.32	1.05 " 1.35	1.00 " 1.40	.90 " 1.50	.60 " 1.80	0 " 2.40
1.30	.5	1.29 " 1.31	1.29 " 1.31	1.28 " 1.32	1.28 " 1.32	1.27 " 1.33	1.24 " 1.36	1.17 " 1.43
	1	1.28 " 1.32	1.27 " 1.33	1.27 " 1.33	1.26 " 1.34	1.24 " 1.36	1.17 " 1.43	1.04 " 1.56
	2	1.26 " 1.34	1.25 " 1.35	1.24 " 1.36	1.21 " 1.39	1.17 " 1.43	1.04 " 1.56	.78 " 1.82
	5	1.19 " 1.41	1.17 " 1.43	1.14 " 1.46	1.08 " 1.52	.97 " 1.62	.65 " 1.95	0 " 2.60
1.40	.5	1.39 " 1.41	1.39 " 1.41	1.38 " 1.42	1.38 " 1.42	1.36 " 1.44	1.33 " 1.47	1.26 " 1.54
	1	1.38 " 1.42	1.37 " 1.43	1.36 " 1.44	1.35 " 1.45	1.33 " 1.47	1.26 " 1.54	1.12 " 1.68
	2	1.35 " 1.45	1.34 " 1.46	1.33 " 1.47	1.31 " 1.49	1.26 " 1.54	1.12 " 1.68	.84 " 1.96
	5	1.28 " 1.52	1.26 " 1.54	1.22 " 1.58	1.17 " 1.63	1.05 " 1.75	.70 " 2.10	0 " 2.80
1.50	.5	1.49 " 1.51	1.48 " 1.52	1.48 " 1.52	1.48 " 1.52	1.46 " 1.54	1.42 " 1.58	1.35 " 1.65
	1	1.48 " 1.52	1.47 " 1.53	1.46 " 1.54	1.45 " 1.55	1.42 " 1.58	1.35 " 1.65	1.20 " 1.80
	2	1.45 " 1.55	1.44 " 1.56	1.42 " 1.58	1.40 " 1.60	1.35 " 1.65	1.20 " 1.80	.90 " 2.10
	5	1.38 " 1.62	1.35 " 1.65	1.31 " 1.69	1.25 " 1.75	1.12 " 1.88	.75 " 2.25	0 " 3.00

A PROPOSAL FOR OBTAINING VARIANCE COMPUTATIONS AS A BY-PRODUCT OF REGULAR TABULATIONS

Margaret Gurney

INTRODUCTION

A common method for computing variances of estimates made from sample data is to divide the sample observations into random groups, and to use the totals of the several random groups as a basis for computation.

A procedure used extensively at the Bureau of the Census permits the use of the standard tabulation programs (written to tabulate the sample) as a device for obtaining the random group totals which are needed to compute variances.

ASSIGNMENT TO RANDOM GROUPS

When the records to be tabulated are edited, weighted, and ready to tabulate, they are put into a sort which reflects the sample design. If multi-stage sampling is used, the whole sample within a first step unit should be assigned to the same random group. The reliability of the estimates of the sampling variance depends on the number of random groups used and on the number of sample elements within each random group. The considerations leading to a choice of the number of random groups are given in HHM, Vol. I.[1] For the purpose of the discussion which follows, however, we use 10 random groups for illustrative purposes.

According to the rules established, each record is assigned to one of the 10 random groups, and the random group number is recorded in an appropriate field of the record. Each of the 10 groups will have approximately one-tenth of the original records in it (some variability will arise if clusters of units are assigned as a whole to a single group). Each of the 10 random groups will, indeed, be a miniature of the whole sample, and the random group totals, when multiplied by 10 will produce estimates similar to those to be obtained from the whole sample. The variance between the 10 random group estimates can be used to represent the variance of the estimate from the whole sample.

TABULATION

The records are sorted for a particular tabulation, with an additional sort on random group as the most minor sort. For example, for a tabulation of farm workers, a tabulation by region, type of farm, economic class of farm, sex and age of the farm operator may be

[1] Hansen, Hurwitz and Madow, Sample Survey Methods and Theory, Vol. I., p 441.

desired. If the tally matrix provides cells for tallying economic class of farm, sex of operator, and age of operator, the only sorts needed for the tabulation of the sample would be by region and type of farm. To obtain random group totals, an additional minor sort by random group is required, so that the sort would be

region
  type of farm
    random group

The tallying puts the records in the cells corresponding to economic class of farm, sex, and age, in the conventional way, throughout the first random group. When the last record of random group 1 has been tallied, the tally matrix is written out on an auxiliary "Random Group Tape." The tally matrix is not cleared at this point.

The records in random group 2 are then tallied in the same matrix, and the cumulative total of random groups 1 and 2 is written out on the Random Group Tape; the tally matrix is not cleared.

The procedure is continued, with a write-out at the end of each random group, until the Random Group Tape has the final write-out, which is the cumulative sum of the 10 random groups for the tally matrix. At this time the tally matrix contains the table which is desired for publication of estimates from the whole sample; this can be written on another tape, and prepared for printing, in the usual way.

UNSCRAMBLING OF "RANDOM GROUP TAPE"

The Random Group Tape will, of course need to have sufficient identification, so that the cells in each write-out of the tally matrix can be identified. It is then a simple matter to obtain the totals for a single random group by subtraction: for example, random group 7 is obtained by subtracting the cumulative totals for groups 1 through 6 from those for groups 1 through 7.

During the "unscrambling" process, a selection of items may be made, if it is not necessary to compute variances for every cell of every table. Moreover, data in a suitable combination of cells may be added, within each random group, and variances may be computed for subtotals and totals.

VARIANCE COMPUTATION

The computation of variances using random group totals is well-known procedure; reference may be made to HHM, Vol I., page 444. If the estimation procedure involves the use of ratio adjustments or composite estimators, these procedures should be applied to each random group separately, before the variance is computed. Otherwise the effect of these features of the estimation procedure will not be reflected in the variance computation.

ADVANTAGES OF THE PROPOSAL

Two major advantages of the proposal described above are:

1. It does not (appreciably) affect the running time of the tabulations required for publication.
2. It does not require the preparation of special tabulation programs, or an early selection of items for which variances are desired.

The additional work involved in getting the random group code into the record is nominal, and may be done as part of some edit, or preliminary (test) tabulation run.

The sort by random group is a minor sort, and follows all other characteristics in the sort required for any particular tally matrix. Hence it requires very little additional in the way of programming and running time.

THE DIFFERENCE BETWEEN THE MEDIAN OF A DIFFERENCE AND THE DIFFERENCE OF THE MEDIANS

Nigel F. Nettheim

THE PROBLEM

Consider a population or sample, finite or infinite, on which two measurements are defined:

$$X \equiv X_1, X_2, \ldots \text{ and } Y \equiv Y_1, Y_2, \ldots$$

Write $X-Y \equiv X_1-Y_1, X_2-Y_2, \ldots$ and write med(X) for the median of X; then it is desired to study the relationship between med(X-Y) and med(X)-med(Y).

This problem arose in studying human life cycles; for example, Y may be the age at first marriage and X may be the age of a mother at the birth of her first child; then if med(X) and med(Y) are available it may be desired to make inferences about med(X-Y), that is, the median time interval between first marriage and birth of first child.

The difference between the medians could in some cases be a perfectly good estimate of the median of the differences; for example, let

$$X \equiv 4, 6, 8 \text{ and } Y \equiv 1, 2, 3;$$

then $X-Y \equiv 3, 4, 5$ and we have med(X-Y) = 4 as well as med(X)-med(Y) = 6-2 = 4.

In other cases the difference between the medians could be a very poor estimate; for example, let

$$X \equiv 150, 200, 101 \text{ and } Y \equiv 1, 50, 100;$$

then $X-Y \equiv 149, 150, 1$ and we have med (X-Y) = 149 by contrast with med(X)-med(Y) = 150-50 = 100.

AN APPROXIMATE SOLUTION

If the random variables X, Y were normally distributed, then X-Y would also be normal, and we would have med(X-Y) = med(X) - med(Y). It is desired to consider the case of non-normality, especially when X and Y are not independent. In the demographic example (which is our motivation), a key feature is the skewness (standardized third moment) of the distributions; this can be introduced by using the Gram-Charlier series expansion for frequency functions (see e.g. Kendall and Stuart Vol. I, p. 156) up to the third order term. It is true that the fourth and higher moments could also differ from those of the normal distribution, and that some of these moments could be taken into account by an extension of the method to be used below but, since this would greatly increase the complexity of the resulting formulae and the

number of parameters to be estimated, it should be hoped that a sufficient indication of the behavior of the estimates will be achieved without referring to such higher moments.

It is shown by J.B.S. Haldane (1942, p. 296) that, to the order of approximation just mentioned,

$$\text{med}(X) \doteqdot \mu_1'(X) - \frac{1}{6}\,\mu_3(X)/\mu_2(X) \qquad (1)$$

where $\mu_1'(X)$ is the mean of X, and $\mu_r(X) \equiv E(X-\mu_1'(X))^r$ is the rth central moment of X. Then also

$$\text{med}(Y) \doteqdot \mu_1'(Y) - \frac{1}{6}\,\mu_3(Y)/\mu_2(Y) \qquad (2)$$

and

$$\text{med}(X-Y) \doteqdot \mu_1'(X-Y) - \frac{1}{6}\,\mu_3(X-Y)/\mu_2(X-Y). \qquad (3)$$

Using the relations

$$\mu_1'(X-Y) = \mu_1'(X) - \mu_1'(Y) \qquad (4)$$

$$\mu_2(X-Y) = \mu_2(X) + \mu_2(Y) - 2\rho(X,Y)\sqrt{\mu_2(X)\mu_2(Y)} \qquad (5)$$

$$\mu_3(X-Y) = \mu_3(X) - \mu_3(Y) + 3\rho(X,Y^2)\sqrt{\mu_2(X)\mu_2(Y^2)} - 3\rho(X^2,Y)\sqrt{\mu_2(X^2)\mu_2(Y)} + 6\rho(X,Y)\sqrt{\mu_2(X)\,\mu_2(Y)}\left[\mu_1'(X) - \mu_1'(Y)\right] \qquad (6)$$

where $\rho(X,Y)$ is the correlation between X and Y, we find the approximate relationship

$$\text{med}(X-Y) \doteqdot \text{med}(X) - \text{med}(Y) + \frac{1}{6}\,\mu_3(X)/\mu_2(X) - \frac{1}{6}\,\mu_3(Y)/\mu_2(Y) - \frac{1}{6}\left\{\mu_3(X) - \mu_3(Y) + 3\rho(X,Y^2)\sqrt{\mu_2(X)\mu_2(Y^2)} - 3\rho(X^2,Y)\sqrt{\mu_2(X^2)\mu_2(Y)} + 6\rho(X,Y)\sqrt{\mu_2(X)\mu_2(Y)}\left[\text{med}(X) - \text{med}(Y) + \frac{1}{6}\mu_3(X)/\mu_2(X) - \frac{1}{6}\mu_3(Y)/\mu_2(Y)\right]\right\} \Big/ \left\{\mu_2(X) + \mu_2(Y) - 2\rho(X,Y)\sqrt{\mu_2(X)\mu_2(Y)}\right\}. \qquad (7)$$

SPECIAL CASES

In general the result (7) involves not only the skewness $(\mu_3/\mu_2^{3/2})$ of the two distributions but also their variances, and it involves not only the correlation between the variables but also the correlation between each variable and the square of the other; it may also be noted that $\mu_2(X^2) \equiv \mu_4(X) - \mu_2^2(X)$. It may be worthwhile giving the simplified formulae which arise in several special cases, as follows.

(i) If X and Y are independent (which will not be the case in the demographic application mentioned above) we have

$$\text{med}(X-Y) \doteqdot \text{med}(X) - \text{med}(Y) + \frac{1}{6}\left[\frac{\mu_3(X)}{\mu_2(X)} - \frac{\mu_3(Y)}{\mu_2(Y)} - \frac{\mu_3(X) - \mu_3(Y)}{\mu_2(X) + \mu_2(Y)}\right]. \qquad (8)$$

(ii) If X and Y are independent and also $\mu_2(X) = \mu_2(Y) = 1$ then

$$\text{med}(X-Y) \doteqdot \text{med}(X) - \text{med}(Y) + \frac{1}{12}(\mu_3(X) - \mu_3(Y)). \qquad (9)$$

(iii) If $\mu_2(X) = \mu_2(Y) = 1$ then

$$\text{med}(X-Y) \doteq [\text{med}(X)-\text{med}(Y)][1-\tfrac{1}{2}\rho(X,Y)/(1-\rho(X,Y))] + \tfrac{1}{6}[\mu_3(X)-\mu_3(Y)] - \tfrac{1}{12}\{[\mu_3(X)-\mu_3(Y)][1+\rho(X,Y)] + 3\rho(X,Y^2)\sqrt{\mu_2(Y^2)} - 3\rho(X^2,Y)\sqrt{\mu_2(X^2)}\}/\{1-\rho(X,Y)\} \quad (10)$$

(iv) If $\mu_3(X) = \mu_3(Y)$ and $\mu_2(X) = \mu_2(Y) = 1$ then

$$\text{med}(X-Y) \doteq [\text{med}(X)-\text{med}(Y)]\{1-\tfrac{1}{2}\rho(X,Y)/[1-\rho(X,Y)]\} - \tfrac{1}{4}\{\rho(X,Y^2)\sqrt{\mu_2(Y^2)} - \rho(X^2,Y)\sqrt{\mu_2(X^2)}\}/\{1-\rho(X,Y)\} \quad (11)$$

ESTIMATION OF PARAMETERS

In order to make use of the result (7), or one of the forms (8) - (11), it is necessary to have some knowledge about the moments, both simple and joint, which appear in those formulae. This information may be obtainable from one of the following sources. (In case none of these methods can be used, it would apparently have to be admitted that the information available is not sufficient to answer the question originally posed.)

(a) Estimate all moments from a sample of the original data, (X,Y), or from a joint distribution thought to be similar to the original one.

(b) Estimate the simple moments from the sample quantiles, if these are available (see for example Benson (1949)).

(c) Make reasonable guesses of the needed moments.

REFERENCES

[1] Benson, F. (1949). A note on the estimation of mean and standard deviation from quantiles. Journal of the Royal Statistical Society, Series B, Vol. XI, No. 1, pp. 91-100.

[2] Haldane, J.B.S. (1942). The mode and median of a nearly normal distribution with given cumulants. Biometrika, Vol. XXXII, pp. 294-299.

[3] Kendall, M.G. and Stuart, A. (1958). The Advanced Theory of Statistics, Vol. 1, Griffin.

THE APPLICATION OF A LINKAGE MODEL TO THE CHANDRASEKAR-DEMING TECHNIQUE FOR ESTIMATING VITAL EVENTS

Benjamin J. Tepping

The Chandrasekar-Deming technique [1] for estimating the number of vital events depends upon determining the number of events reported by both of two sources (for example, a sample survey and an official registration system). Thus, the matching of events reported on two lists is critical. Jabine and Bershad [2], Marks [3] and Seltzer and Adlakha [4] have discussed the effect of matching errors on the bias of the Chandrasekar-Deming estimator. The present writer [5] has presented a general model for the linkage of records. This note is an attempt to apply that model to the problem of matching in the Chandrasekar-Deming technique, and especially to suggest an appropriate utility function as required by the model. It is assumed in the sequel that the reader is familiar with [5].

Assume that for any pair of events the possible actions are:

a_1: link

a_2: nonlink

a_3: field check of the pair, resulting in treatment of the pair as link or nonlink, respectively sub-actions a_{31} or a_{32}.

We assume that a_{31} and a_{32} are stochastic, so that we may speak of $P(a_{31}|M,\gamma)$ $P(a_{32}|M,\gamma)$, i.e., the conditional probabilities given that a pair subjected to a field check is truly a match and has the comparison vector γ, that the pair will be treated as a link or a nonlink. Let the cost of conducting the field check for a single pair be K. Then the utilities of the several actions may be written:

$$U(a_1|\gamma) = U(a_1|M)P(M|\gamma) + U(a_1|\bar{M})P(\bar{M}|\gamma)$$

$$U(a_2|\gamma) = U(a_2|M)P(M|\gamma) + U(a_2|\bar{M})P(\bar{M}|\gamma)$$

$$U(a_3|\gamma) = U(a_3|M)P(M|\gamma) + U(a_3|\bar{M})P(\bar{M}|\gamma)$$

Further,

$$U(a_3|M)=U(a_1|M)P(a_{31}|M,\gamma)+U(a_2|M)P(a_{32}|M,\gamma)-K$$

$$U(a_3|\bar{M})=U(a_1|\bar{M})P(a_{31}|\bar{M},\gamma)+U(a_2|\bar{M})P(a_{32}|\bar{M},\gamma)-K$$

We note that

$$P(M|\gamma) = 1 - P(\bar{M}|\gamma) = x(\gamma), \text{ say}$$

$$P(a_{31}|M,\gamma) = 1 - P(a_{32}|M,\gamma) = y(\gamma), \text{ say}$$

$$P(a_{32}|\bar{M},\gamma) = 1 - P(a_{31}|\bar{M},\gamma) = z(\gamma), \text{ say.}$$

We then have

$$U(a_1|\gamma) = U(a_1|M)x + U(a_1|\bar{M})(1-x)$$

$$U(a_2|\gamma) = U(a_2|M)x + U(a_2|\bar{M})(1-x)$$

$$U(a_3|\gamma) = \left[U(a_1|M)y + U(a_2|M)(1-y) - K\right]x + \left[U(a_1|M)(1-z) + U(a_2|\bar{M})z-K\right](1-x)$$

Simplifying somewhat, we get

$$U(a_1|\gamma)=\left[U(a_1|M)-U(a_1|\bar{M})\right]x+U(a_1|\bar{M})$$

$$U(a_2|\gamma)=\left[U(a_2|M)-U(a_2|\bar{M})\right]x+U(a_2|\bar{M})$$

$$U(a_3|\gamma)=\left[\left\{U(a_1|M)-U(a_2|M)\right\}y+\left\{U(a_1|\bar{M})-U(a_2|\bar{M})\right\}z\right.$$
$$\left.+U(a_2|M)-U(a_1|\bar{M})\right]x$$
$$+\left[U(a_2|\bar{M})-U(a_1|\bar{M})\right]z+U(a_1|\bar{M})-K$$

Let us denote

$$C_{11}=U(a_1|M) \qquad C_{12}=U(a_1|\bar{M})$$
$$C_{21}=U(a_2|M) \qquad C_{22}=U(a_2|\bar{M})$$

Then

$$U(a_1|\gamma)=(C_{11}-C_{12})x+C_{12}$$
$$U(a_2|\gamma)=(C_{21}-C_{22})x+C_{22}$$
$$U(a_3|\gamma)=\left[(C_{11}-C_{21})y+(C_{12}-C_{22})z+C_{21}-C_{12}\right]x$$
$$-(C_{12}-C_{22})z+C_{12}-K$$

The utility surfaces for a_1 and a_2 are hyperplanes over the (x,y,z)-space but the utility surface for a_3 is hyperbolic. The unit cube in the (x,y,z)-space is divided into 3 decision regions according as $U(a_1|\gamma)$, $U(a_2|\gamma)$ or $U(a_3|\gamma)$ is the greatest. Now

$$U(a_1|\gamma)\gtrless U(a_2|\gamma) \text{ iff } (C_{11}-C_{12}-C_{21}+C_{22})x+C_{12}-C_{22}\gtrless 0$$

$$U(a_1|\gamma)\gtrless U(a_3|\gamma) \text{ iff } \left[C_{11}-C_{21}-(C_{11}-C_{21})y-(C_{12}-C_{22})z\right]x+(C_{12}-C_{22})z+K\gtrless 0$$

$$U(a_2|\gamma)\gtrless U(a_3|\gamma) \text{ iff } \left[C_{12}-C_{22}-(C_{11}-C_{21})y-(C_{12}-C_{22})z\right]x+(C_{12}-C_{22})z+C_{22}-C_{12}+K\gtrless 0$$

From these inequalities we can calculate the decision regions in the unit cube for specified values of C_{ij}.

Our problem now is to specify the values of the utilities C_{ij} and the cost K of making a field check for a single pair of events. Without loss of generality we may take K=1, thus specifying the C_{ij} in units of the cost of a field check.

Let us assume that the matching rule adopted links each event contained in the sample with no more than one event contained in the register. Let a' denote the number of sample events linked with the register, and b' the number of sample events not linked with the register. If the sample is self-weighting, the estimated number of matching events is ka' and the estimated number of non-matching events (i.e., the number of events discoverable by the sample survey process, but not contained in the register) is kb'. Let R denote the total number of events contained in the register, so that the CD estimate of total events in the population may be written

$$x' = \frac{R(a'+b')}{a'} = \frac{Rn}{a'}$$

where $n = a'+b'$ denotes the number of sample events. We may write

$$a' = n_{11} + n_{121} + n_{122}$$

$$b' = n_{21} + n_{22}$$

where

n_{11} = number of sample events linked with a matching event of the register

n_{121} = number of sample events linked with a non-matching event of the register, but a matching event exists in the register

n_{122} = number of sample events linked with a non-matching event of the register, and no matching event exists in the register

n_{21} = number of sample events not linked with any event of the register, but a matching event exists in the register

n_{22} = number of sample events not linked with any event of the register, and no matching event exists in the register.

Now,

$$x' = \frac{Rn}{n_{11}+n_{121}+n_{122}} .$$

If the true matches and nonmatches in the sample were known, the CD estimate would be

$$x'' = \frac{Rn}{n_{11}+n_{121}+n_{21}}$$

so that the error of x' resulting from linkage errors is

$$x'-x'' = Rn \frac{n_{21}-n_{122}}{(n_{11}+n_{121}+n_{122})(n_{11}+n_{121}+n_{21})}$$

and the relative error of x' resulting from linkage errors is

$$\epsilon = \frac{x'-x''}{x''} = \frac{n_{21}-n_{122}}{n_{11}+n_{121}+n_{21}} .$$

It may be noted that ϵ is zero if $n_{21} = n_{122}$, as would be expected.

For a fixed value of n we may replace one of the five variables by a function of the other four, say

$$n_{22} = f(n_{11}, n_{121}, n_{122}, n_{21})$$

and then write

$$\frac{\partial \epsilon}{\partial n_{11}} = -\frac{n_{21}-n_{122}}{(n_{11}+n_{121}+n_{21})^2}$$

$$\frac{\partial \epsilon}{\partial n_{121}} = -\frac{n_{21}-n_{122}}{(n_{11}+n_{121}+n_{21})^2}$$

$$\frac{\partial \epsilon}{\partial n_{122}} = -\frac{1}{n_{11}+n_{121}+n_{21}}$$

$$\frac{\partial \epsilon}{\partial n_{21}} = -\frac{n_{11}+n_{121}+n_{122}}{(n_{11}+n_{121}+n_{21})^2}$$

These expressions for the derivatives suggest that the rate of change of ϵ with respect to n_{11} and n_{121} are negligible, and that (except for the direction of the error) the rate of change of ϵ with respect to n_{122} and n_{21} are roughly the same, for

$$\frac{\partial \epsilon}{\partial n_{21}} = -\frac{\partial \epsilon}{\partial n_{122}} \left[1 - \frac{n_{21}-n_{122}}{n_{11}+n_{121}+n_{21}}\right]$$

If we take negative and positive errors to be equally important, we may take

$$C_{11} = C_{22} = 0$$

$$C_{12} = C_{21} = -V$$

where V is the amount that one judges he should be willing to pay (in units of the cost of a field check for one event) for a reduction of $\frac{1}{m}$ in the relative error of the estimate, m being the number of sample events which are also contained in the register.

We may now determine the decision regions by the conditions

$$U(a_1|\gamma) \gtrless U(a_2|\gamma) \quad \text{iff} \quad x \gtrless 1/2$$

$$U(a_1|\gamma) \gtrless U(a_3|\gamma) \quad \text{iff} \quad x \gtrless \frac{z-1/V}{1-y+z}$$

$$U(a_2|\gamma) \gtrless U(a_3|\gamma) \quad \text{iff} \quad x \lesseqgtr \frac{1-z+1/V}{1+y-z}$$

Thus the actions a_1, a_2, a_3 are indicated by the conditions

a_1: $x > \max\left[1/2, \frac{z-1/V}{1-y+z}\right]$

a_2: $x < \min\left[1/2, \frac{1-z+1/V}{1+y-z}\right]$

a_3: all other values of x.

The contours of the surfaces which bound the decision regions are shown in the graphs for selected values of V. For example, if V=20, y=.8, and z=.9,

the pair should be a link if $x > .77$

the pair should be a nonlink if $x < .17$

the pair should be field-checked if

$$.17 \leqq x \leqq .77.$$

It is of interest to note that the decision regions are only mildly sensitive to the value V. Thus if V=40 in the example above, a field check would take place if $.14 \leqq x \leqq .80$, while if V=10 the interval would be $.22 \leqq x \leqq .73$.

REFERENCES

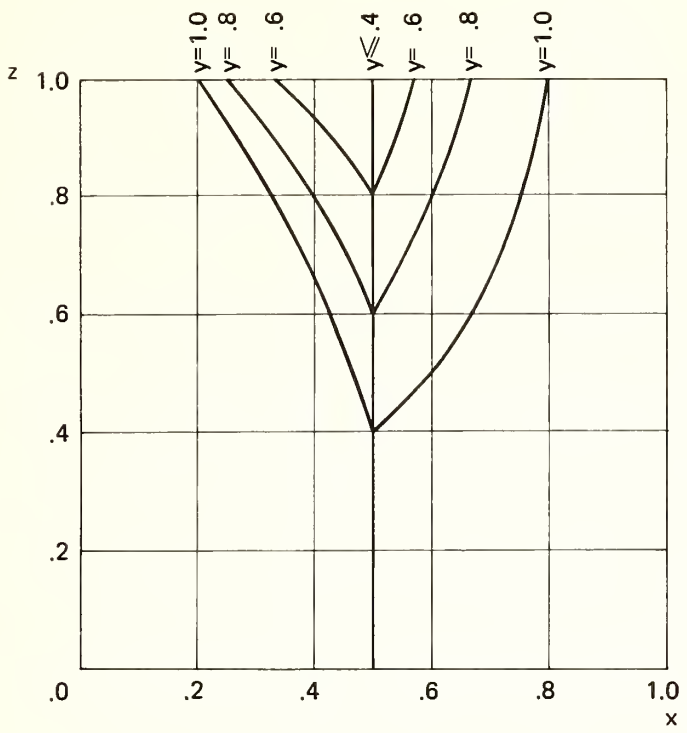
[1] C. Chandra Sekar and W. Edwards Deming, "On a method of estimating birth and death rates and the extent of registration," J.A.S.A., 44 (1949), pp. 101-115.

[2] Thomas B. Jabine and Max A. Bershad, "Some comments on the Chandrasekeran-Deming technique for the measurement of population change," unpublished manuscript presented at CENTO symposium on demographic statistics, Karachi, Pakistan, October 30, 1968.

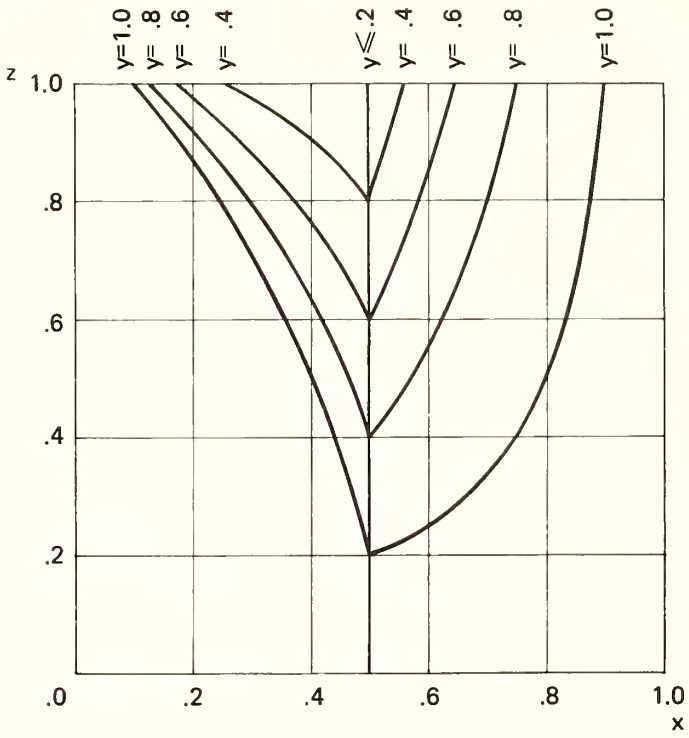
[3] Eli S. Marks, "Some comments on the Chandrasekar-Deming estimation technique," unpublished manuscript, January 20, 1969.

[4] William Seltzer and Arjun Adlakha, "On the effect of errors in the application of the Chandrasekar-Deming technique," unpublished manuscript, January 21, 1969.

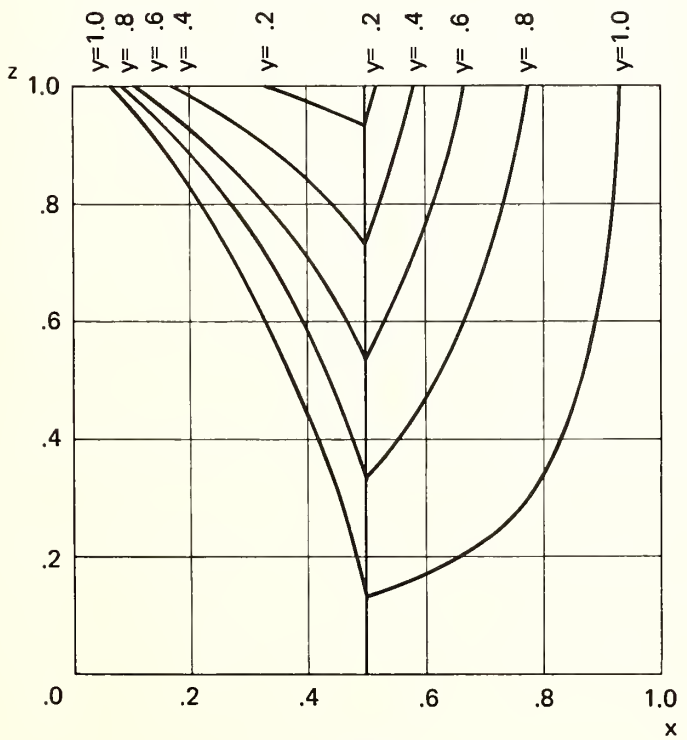
[5] Benjamin J. Tepping, "A model for optimum linkage of records," J.A.S.A., 63 (1968), pp. 1321-1332.



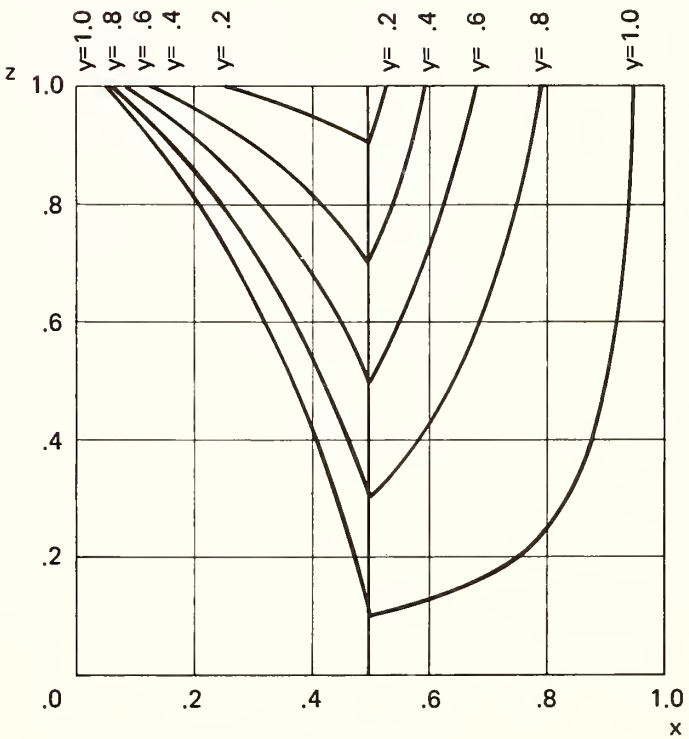
V = 5



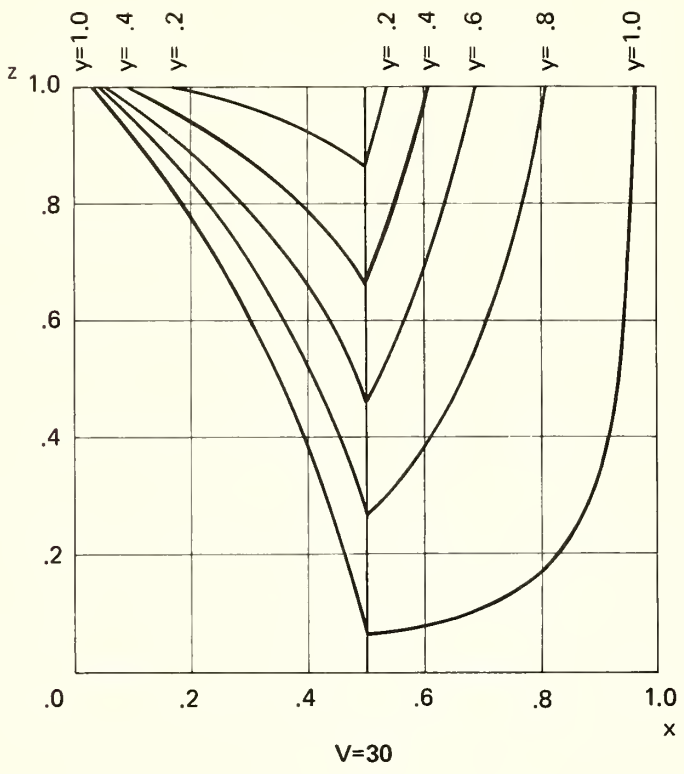
V=10



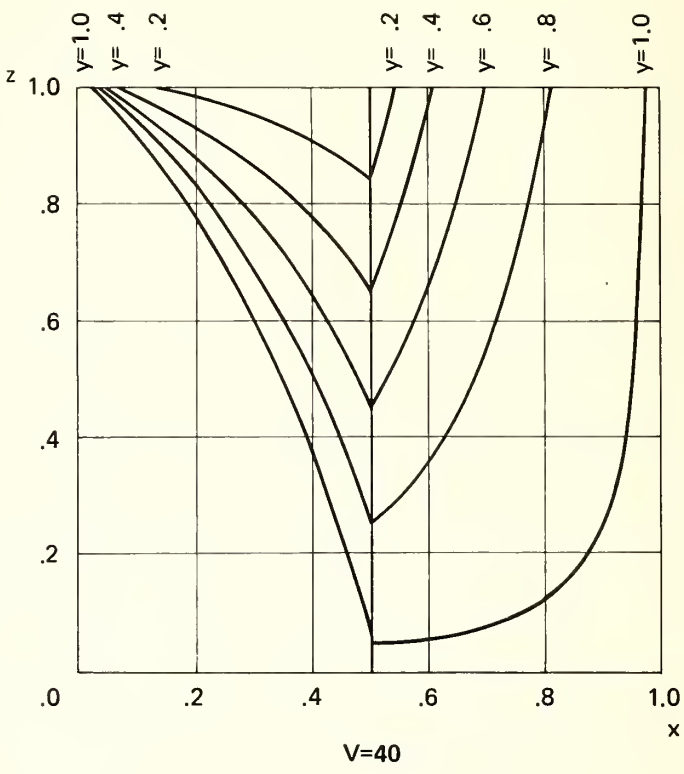
V=15



V=20


y=1.0
y= .4
y= .2
y= .2
y= .4
y= .6
y= .8
y=1.0
z
1.0
.8
.6
.4
.2
.0
.2
.4
.6
.8
1.0
x

V=30


y=1.0
y= .4
y= .2
y= .2
y= .4
y= .6
y= .8
y=1.0
z
1.0
.8
.6
.4
.2
.0
.2
.4
.6
.8
1.0
x

V=40

THE USE OF THE PLOTTING MACHINE IN PLOTTING SCATTER DIAGRAMS

Benjamin J. Tepping

The plotting machine associated with the IBM 1401 computer in the Bureau of the Census is sometimes used to plot "scatter diagrams," i.e., a set of discrete points (x_i, y_i). The points are presented to the machine in a specified order, and the machine plots the $(i + 1)$-th point by moving the plotting pen a distance $|x_{i+1} - x_i|$ horizontally and then a distance $|y_{i+1} - y_i|$ vertically. Thus, the total distance moved in the course of plotting n points is

$$\sum_{i=1}^{n-1} \left\{ |x_{i+1} - x_i| + |y_{i+1} - y_i| \right\}.$$

The purpose of this note is to examine the question of how to arrange the data for presentation to the plotter so as to minimize the total distance, and therefore the time used on the plotter.

RECOMMENDATION: Sort the points into S "strips", divide the range of y into S equal intervals and classify the points into these intervals according to the value of y_i. Take for S an integer which approximates

$$S = \sqrt{\frac{n\ b}{3\ a}}$$

where b is the total range of y, a is the average range of x within a strip, and n is the number of points to be plotted. Within each strip, sequence the points according to the value of x_i, ascending and descending in alternate strips.

Example:

$$n = 120$$
$$b = 5$$
$$a = 1$$
$$S \doteq \sqrt{200} \doteq 14$$

The saving that can be accomplished is a function primarily of the number of points, n, to be plotted. The following table illustrates the saving, for the case a = b, in the distance travelled to traverse n points selected at random from a uniform distribution.

n	Saving
100	83 %
200	88 %
500	92.2%
1000	94.5%
2000	96.1%
5000	97.6%

DISCUSSION

It is evident that the length of a path given by any prescribed rule will depend upon the distribution of the points to be plotted. We assume here that these points are a random sample of n from a uniform distribution over the rectangle

$$0 < x \leq a, \qquad 0 < y \leq b.$$

Case 1: The sequence of points (x_i, y_i)

is in random order. The total length of path is then

$$\ell = \sum_{i=1}^{n-1} \left\{ |x_i - x_{i+1}| + |y_i - y_{i+1}| \right\}$$

But since x_i and x_{i+1} are a random sample of 2 from a uniform distribution in the interval $0 < x \leq a$,

$$E|x_i - x_{i+1}| = \frac{a}{3}.$$

Similarly,

$$E|y_i - y_{i+1}| = \frac{b}{3}.$$

Hence

$$E\ \ell = \frac{n-1}{3}(a + b).$$

Case 2: The sequence of points (x_i, y_i) is such that the x_i are the order statistics:

$$x_1 \leq x_2 \leq \cdots \leq x_n.$$

Then, again,

$$E\ |y_i - y_{i+1}| = \frac{b}{3}$$

but

$$E\ |x_i - x_{i+1}| = \frac{a}{n+1}\ .$$

Hence

$$E\ \ell = (n-1)\left[\frac{a}{n+1} + \frac{b}{3}\right]$$

Case 3: The points (x_i, y_i) are classified into S strips, each of width b/S. The point (x_i, y_i) is in the j-th strip if

$$\frac{(j-1)b}{S} < y_i \leq \frac{jb}{S}\ .$$

Within each strip, the points are ordered according to the value of x_i, ascending and descending in alternate strips. This defines the sequence in which the points are traversed. Let n_j be the number of points in the j-th strip, and let (x_{jk}, y_{jk}) denote the k-th point in the j-th strip. Then the total length of path is

$$\ell = \sum_{j=1}^{S} \sum_{k=1}^{n_j-1} \left\{ |x_{j,k} - x_{j,k+1}| + |y_{j,k} - y_{j,k+1}| \right\}$$

$$+ \sum_{j=1}^{S-1} \left\{ |x_{j,n_j} - x_{j+1,1}| + |y_{j,n_j} - y_{j+1,1}| \right\}.$$

As before,

$$E\left[(|x_{j,k} - x_{j,k+1}|) \,\middle|\, n_j \right] = \frac{a}{n_j+1}$$

$$E\left[(|y_{j,k} - y_{j,k+1}|) \,\middle|\, n_j \right] = \frac{b}{3S}\ .$$

Moreover,

$$E\left[(|y_{j,n_j} - y_{j+1,1}|) \,\middle|\, n_j \right] = \frac{b}{S}\ .$$

For the moment let us define

$$d_j = E\left[(|x_{j,n_j} - x_{j+1,1}|) \,\middle|\, n_j, n_{j+1} \right]$$

Then

$$E\left[\ell \,|\, (n_j) \right] = \sum_{j=1}^{S} \sum_{k=1}^{n_j-1} \left\{ \frac{a}{n_j+1} + \frac{b}{3S} \right\} + \sum_{j=1}^{S-1} \left\{ \frac{b}{S} + d_j \right\}$$

$$= \sum_{j=1}^{S} \left\{ \frac{n_j-1}{n_j+1}\, a + \frac{(n_j-1)b}{3S} \right\} + \frac{S-1}{S}\, b + \sum_{j=1}^{S-1} d_j$$

$$= a \sum_{j=1}^{S} \frac{n_j-1}{n_j+1} + \frac{b}{3S}(n + 2S - 3) + \sum_{j=1}^{S-1} d_j.$$

Hence

$$E\ \ell = a \sum_{j=1}^{S} E\, \frac{n_j-1}{n_j+1} + \frac{b}{3S}(n + 2S - 3) + \sum_{j=1}^{S-1} E\, d_j$$

$$= a\, S\, E\, \frac{n_j-1}{n_j+1} + \frac{b}{3S}(n + 2S - 3) + (S-1)\, E\, d_j.$$

Since n_j is binomially distributed with parameters n and $\frac{1}{S}$, it can be shown that

$$E\,\frac{n_j-1}{n_j+1} = 1 - \frac{2S}{n+1}\left[1 - \left(\frac{S-1}{S}\right)^{n+1}\right]$$

$$\doteq 1 - \frac{2S}{n+1}\left[1 - e^{\frac{-n+1}{S}}\right]$$

$$\doteq 1 - \frac{2S}{n+1}$$

To evaluate d_j, we may without loss of generality take x_{j,n_j} and $x_{j+1,1}$ to be the maxima in the j-th and (j+1)-th strips respectively. It can be shown (See Appendix A) that

$$d_j = \frac{a}{n_j+n_{j+1}+1}\left(\frac{n_j}{1+n_{j+1}} + \frac{n_{j+1}}{1+n_j}\right).$$

Then for reasonably large values of n, an approximation is

$$E\,d_j \doteq \frac{a}{\frac{2n}{S}+1} \times 2\,\frac{\frac{n}{S}}{1+\frac{n}{S}} = \frac{2a}{(\frac{2}{S}+\frac{1}{n})(S+n)}$$

$$\doteq \frac{a}{1+\frac{n}{S}} \doteq \frac{aS}{n}.$$

Thus

$$E\ell = aS\left(1 - \frac{2S}{n+1}\right) + \frac{b}{3S}(n + 2S - 3) + \frac{aS(S-1)}{n}.$$

Now to find the value of S which minimizes $E\ell$, we set

$$0 = \frac{d}{dS}\,E\ell = a - \frac{4\,aS}{n+1} - \frac{bn}{3S^2} + \frac{b}{S^2} + \frac{a(2S-1)}{n}$$

$$= a\left[\frac{n-1}{n} - \frac{2(n-1)}{n(n+1)}\,S - \frac{b}{a}\,\frac{n-3}{3}\,\frac{1}{S^2}\right]$$

$$= -ak\,\frac{n-1}{n}\,\frac{1}{S^2}\left[\frac{2}{n+1}\,\frac{1}{k}\,S^3 - \frac{S^2}{k} + \frac{n}{n-1}\right]$$

where

$$k = \frac{b}{a}\,\frac{n-3}{3}.$$

Thus S is a solution of the cubic equation

$$\frac{2}{k(n+1)}S^3 - \frac{S^2}{k} + \frac{n}{n-1} = 0,$$

which depends on n and the ratio b/a, or

$$\frac{2}{n+1}\,S^3 - S^2 + \frac{n}{n-1}\,k = 0.$$

We observe that

$$\frac{d^2}{dS^2}\,E\ell = a\left[-\frac{2(n-1)}{n(n+1)} + \frac{2K}{S^3}\right]$$

which is positive if and only if

$$S^3 < \frac{b}{a}\,\frac{n(n+1)(n-3)}{3(n-1)}$$

But the derivative of the left-hand side of the cubic equation to be solved is $2S(\frac{3S}{n+1} - 1)$, so that one of the two positive solutions of the cubic equation is less than (n+1)/3 and the other is greater than (n+1)/3. For values of b/a and n to be used in practice,

$$(\tfrac{n+1}{3})^3 > \frac{b}{a}\,\frac{n(n+1)(n-3)}{3(n-1)}$$

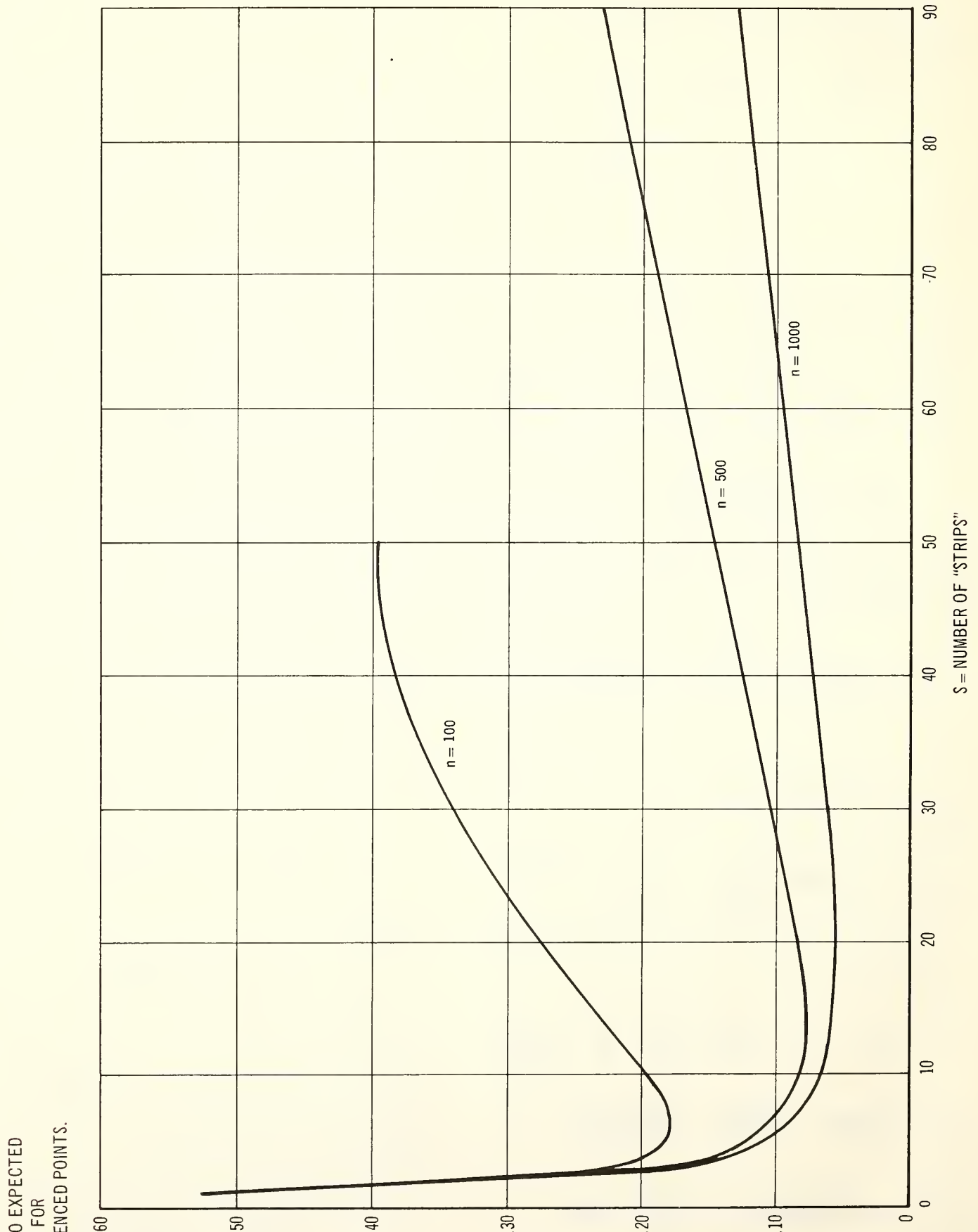
since

$$(n+1)^2 > \frac{b}{a}\,\frac{9n(n-3)}{n-1}$$

and certainly b/a < 5 so that the inequality will be satisfied when n > 45. Hence the minimum sought is the smaller of the two positive roots of the cubic.

For the usual values of n, the value

$$S = \sqrt{\frac{n}{n-1}}\,k = \sqrt{\frac{n}{n-1}\,\frac{b}{a}\,\frac{n-3}{3}}$$


RATIO TO EXPECTED
LENGTH FOR
UNSEQUENCED POINTS.
.60
.50
.40
.30
.20
.10
0
0
10
20
30
40
50
60
70
80
90
n = 100
n = 500
n = 1000
S = NUMBER OF "STRIPS"

is a good approximation to the root we seek. (See appendix B for examples.) An easy approximation to use is

$$S = \sqrt{\frac{n}{3}\frac{b}{a}}$$

The accompanying graph illustrates the saving that may be achieved as compared with plotting random, unsequenced points. For large n (say $n \geq 100$), the expected length of path for unsequenced points is approximately $2an/3$, where a is the side of the square within which the population of points is uniformly distributed. If the points are sequenced within s strips, with $s = \sqrt{n/3}$, the expected length of path is approximately $2a\sqrt{n/3}$. The ratio of lengths is therefore approximately $\sqrt{n/3}$.

The graph also illustrates the relative insensitivity of the length to small increases in the value of s that is used.

The discussion above has been for the case of points distributed uniformly over a rectangle. The recommendation with which this note begins is based on the following argument. For a general distribution of points, let the range of y be divided in S_o strips. If S_o is so chosen that

$$S_o = \sqrt{\frac{n}{3}\frac{b}{R_o}}$$

where R_o is the average range of x within a strip, it will usually be found in practice that the points within a strip are much more nearly distributed uniformly over a rectangle than the aggregate of all the points. Within a strip, in such case, the optimum number of substrips in the average strip would be approximately

$$\sqrt{\frac{n/S_o}{3}\frac{b/S_o}{R_o}} = 1$$

This leads to the recommendation.

APPENDIX A

The density function of the maximum in a sample of n from a uniform distribution in (0,1) is $h(u) = n\,u^{n-1}$. Let v be the maximum in an independent sample of size m, so that $g(v) = m\,v^{m-1}$. Let $w = u - v$. Then

$$F(w) = \begin{cases} \int_{v=-w}^{1} \int_{u=o}^{w+v} h(u)g(v)\,du\,dv, & -1 < w \leq 0 \\ 1 - \int_{v=o}^{1-w} \int_{u=w+v}^{1} h(u)g(v)\,du\,dv, & 0 < w \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

Substituting the explicit forms of g(v) and h(u), and integrating with respect to u, we get

$$F(w) = \begin{cases} m \int_{v=-w}^{1} v^{m-1}\,(w+v)^n\,dv, & -1 < w \leq 0 \\ 1 - m \int_{v=o}^{1-w} v^{m-1}\left[1 - (w+v)^n\right] dv, & 0 < w \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

Therefore,

$f(w) = F'(w)$

$$= \begin{cases} m\,n \int_{v=-w}^{1} v^{m-1} (w+v)^{n-1}\, dv, & -1 < w \leq 0 \\ m\,n \int_{v=0}^{1-w} v^{m-1} (w+v)^{n-1}\, dv, & 0 < w \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

Now

$$E|w| = \int_{w=-1}^{0} (-w)\, f(w)\, dw + \int_{w=0}^{1} w\, f(w)\, dw$$

$$\int_{w=0}^{1} wf(w)dw = m\,n \int_{w=0}^{1} w \int_{v=0}^{1-w} v^{m-1} (w+v)^{n-1}\, dv$$

$$= m\,n \int_{w=0}^{1} \int_{v=0}^{1-w} wv^{m-1} \sum_{r=0}^{n-1} \binom{n-1}{r} w^{r} v^{n-1-r}\, dv\, dw$$

$$= m\,n \sum_{r=0}^{n-1} \binom{n-1}{r} \int_{w=0}^{1} \int_{v=0}^{1-w} w^{r+1} v^{m+n-2-r}\, dv\, dw$$

$$= m\,n \sum_{r=0}^{n-1} \binom{n-1}{r} \frac{1}{m+n-1-r} \int_{0}^{1} w^{r+1} (1-w)^{m+n-1-r}\, dw$$

$$= m\,n \sum_{r=0}^{n-1} \binom{n-1}{r} \frac{1}{m+n-1-r} \frac{(r+1)!\,(m+n-1-r)!}{(m+n+1)!}$$

$$= m\,n \sum_{r=0}^{n-1} \frac{(n-1)!}{(n-1-r)!} (r+1) \frac{(m+n-2-r)!}{(m+n+1)!}$$

$$= m\,n \sum_{x=0}^{n-1} \frac{(n-1)!}{x!} (n-x) \frac{(x+m-1)!}{(m+n+1)!}$$

$$= \frac{m\,n!}{(m+n+1)!} \sum_{x=0}^{n-1} (n-x) \frac{(x+m-1)!}{x!}$$

$$= \frac{m\,n!}{(m+n+1)!} \left\{ n \sum_{x=0}^{n-1} \frac{(x+m-1)!}{x!} - \sum_{x=1}^{n-1} \frac{(x+m-1)!}{(x-1)!} \right\}$$

Using the notation of the calculus of finite differences,

$$\sum_{x=0}^{n-1} \frac{(x+m-1)!}{x!} = \sum_{x=0}^{n-1} (x+m-1)_{m-1}$$

$$= \sum_{x=0}^{n-1} \frac{1}{m} \Delta\, (x+m-1)_{m}$$

$$= \frac{1}{m} (n+m-1)_{m} = \frac{1}{m} \frac{(n+m-1)!}{(n-1)!}$$

$$\sum_{x=1}^{n-1} \frac{(x+m-1)!}{(x-1)!} = \sum_{x=1}^{n-1} (x+m-1)_{m}$$

$$= \sum_{x=1}^{n-1} \frac{1}{m+1} \Delta\, (x+m-1)_{m+1}$$

$$= \frac{1}{m+1} (n+m-1)_{m+1} = \frac{1}{m+1} \frac{(n+m-1)!}{(n-2)!}$$

Hence

$$\int_{w=0}^{1} wf(w)\, dw = \frac{m\,n!}{(m+n+1)!} \left\{ \frac{n}{m} \frac{(n+m-1)!}{(n-1)!} - \frac{1}{m+1} \frac{(n+m-1)!}{(n-2)!} \right\}$$

$$= \frac{mn}{(m+n)(m+n+1)} \left\{ \frac{n}{m} - \frac{n-1}{m+1} \right\}$$

$$= \frac{n}{(m+1)(m+n+1)}$$

Similarly, or by a simple argument of symmetry,

$$\int_{w=-1}^{0} (-w) f(w)\, dw = \frac{m}{(n+1)(m+n+1)}$$

Therefore

$$E|w| = \frac{1}{m+n+1} \left(\frac{m}{n+1} + \frac{n}{m+1} \right)$$

Furthermore, if the range of the uniform distribution is $\underline{a}$, it is clear that the expected value is simply multiplied by $\underline{a}$.

APPENDIX B

Examples:

I. Take b/a = 1, n = 100. Then $k = \frac{97}{3}$

and we need to solve the cubic equation

$$\phi(S) = \frac{2}{101} S^3 - S^2 + \frac{100}{99} \times \frac{97}{3} = 0.$$

The approximation suggested is $S_o = 10\sqrt{\frac{97}{3 \times 99}}$

= 5.714887, while the actual root is

S= 6.094453.

II. Take b/a = 2, n = 500. Then $k = \frac{994}{3}$.

$$\phi(S) = \frac{2}{501} S^3 - S^2 + \frac{500}{499} \times \frac{994}{3} = 0.$$

Then

$$S_o = 18.220794 \text{ and } S = 18.951800.$$

III. Take b/a = 1, n = 2000. Then $k = \frac{1997}{3}$.

$$\phi(S) = \frac{2}{2001} S^3 - S^2 + \frac{2000 \times 1997}{3 \times 1999}.$$

Then

$$S_o = 25.806969 \text{ and } S = 26.150999.$$

IV. Take b/a = 5, n = 120. Then

$k = \frac{5 \times 117}{3}$

$$\phi(S) = \frac{2}{121} S - S^2 + \frac{120}{119} \times \frac{5 \times 117}{3}.$$

Then

$$S_o = 14.022791 \text{ and } S = 16.430142.$$

☆ U. S. GOVERNMENT PRINTING OFFICE : 1971 431-995/1366

U.S. DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, D.C. 20233

OFFICIAL BUSINESS

PENN STATE UNIVERSITY LIBRARIES
A000070962253